Design and Analysis of Knowledge Based Explainable Intelligent Systems

Synopsis

Submitted in the partial fulfilment for the requirement of

Doctor of Philosophy

In

COMPUTER SCIENCE

Submitted by

Abhishek Toofani

Under the supervision of

Prof. Sandeep Paul
(Co-Supervisor)
Associate Professor
Dept. of Physics & Computer Science

Dr. Lotika Singh
(Supervisor)
Assistant Professor
Dept. of Physics & Computer Science

Department of Physics And Computer Science, Faculty of Science

Dayalbagh Educational Institute, (Deemed To Be University)

Dayalbagh, Agra, UP, 282005

2022

1. Introduction

Artificial intelligence is a boon for humans. From SIRI, and Alexa to Tesla cars, AI is a part of daily life now. Not only for end users, but AI is also used in industries like manufacturing, financing, medical science, and many more. For example, in Covid-19 we have seen smart devices like tomography thermal scanning in public places to recognize any infection. Yes, it is humans which provide machines with the capabilities of reasoning, learning, planning, and creativity, but the question is, can a human have blind faith in every decision taken by an AI machine. Following are some interesting examples that make us think about this question in deep.

In 2013 Eric Loomis was sentenced to six years of prison which was partly based on COMPAS [1] scale, it was a probabilistic-based algorithm used to predict an individual's risk of recidivism. Eric Loomis appealed against the sentence by objecting to the use of COMPAS because its risk assessment procedure is not transparent to the public. Still, Court sentenced him because of his criminal history.

In one more incident, in the historic Go game match between Lee Sedol, and AlphaGo (an AI system by DeepMind) [2], AlphaGo played a move that was totally unexpected. Even the expert team had no idea how AlphaGo came to this decision. An expert of AlphaGo had to comment on the incident "It's not a human move. I've never seen a human play this move."

These are some examples that bring a worry to humans as well as AI developers about the decision-making ability of AI systems because decisions taken by these machines are not transparent. Due to this non-transparent tendency, developers are not able to explain the decision made by any system to the end user which eventually makes a barrier to safely using new AI technologies. It, therefore, becomes essential for an AI system to be explainable.

In this respect, The UK Parliament House of Lords select committee in their report mentioned that "*We believe it is not acceptable to deploy any artificial intelligence system which could have a substantial impact on an individual's life, unless it can generate a full and satisfactory explanation for the decisions it will take*" [3]. The European Parliament has also recognized this and they recently adopted General Data Protection Regulation (GDPR) as a clause which reiterates the right of all individuals to obtain "*meaningful explanations of the logic involved when automated decision making takes place*" [4].

To make machine learning more understandable, DARPA (USA Defence Advanced Research Projects Agency) announced a new program called Explainable Artificial Intelligence (XAI). This program was called the "third-wave AI systems" [5]. In this program, understanding of models and the working environment increases over time. These models are called explanatory models as phenomena behind these AI machines are open to characterize the behavior of the model.

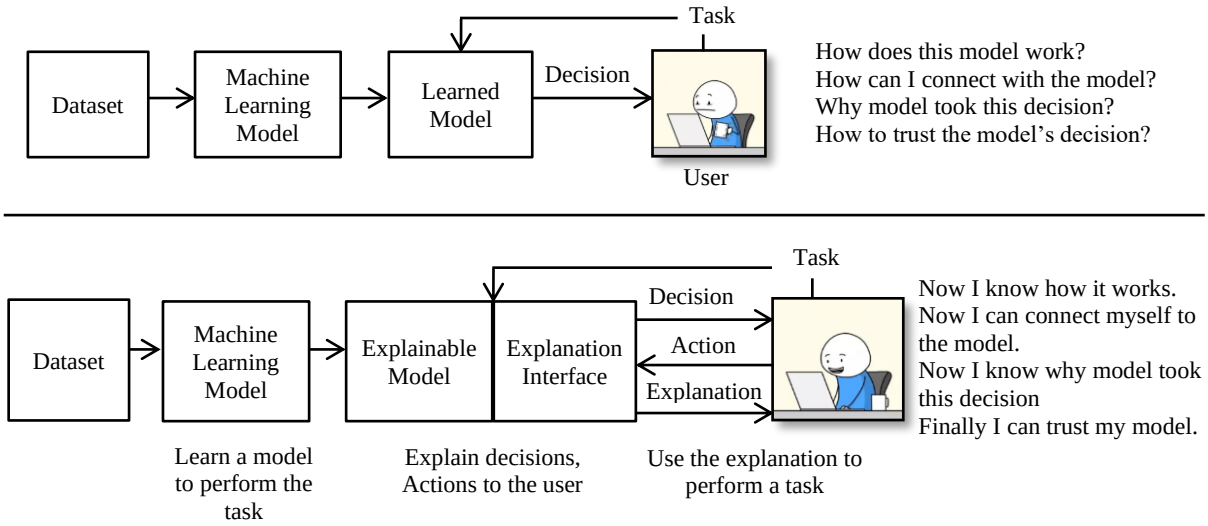

Figure 1: Basic description of XAI

Figure 1 describes the DRPA point of view in XAI. When XAI system takes any action or gives any recommendation, it should provide a full explanation of the decision taken. The user can then make a decision based on the explanation provided by the XAI system. Therefore, the role of XAI is to provide an explanation of the individual decision, enable an understanding of overall strengths and weaknesses, and convey an understanding of how the system will behave in the future and how to correct the system's mistakes.

In machine learning, many algorithms exist which may have very high accuracy, for example, the deep learning model is a high accuracy model however, a serious limitation of these types of models is that they are not able to explain how the output is generated or why machine takes a particular decision. These types of models are called the Black box model (BBM).

Figure 2 shows the relation between accuracy and interpretability of some well-known machine learning algorithms. As shown in the figure Linear/Logistic regression, Decision Tree, or any probabilistic model has good interpretability or explainability but their accuracy is not up to the mark. However, NN, SVM, or deep learning models have high accuracy but have poor interpretability.

Therefore, there is a need to design an interpretable or explainable model which can also provide higher accuracy. The next section shows some latest work on XAI where researchers explain the working of a BBM through various techniques.

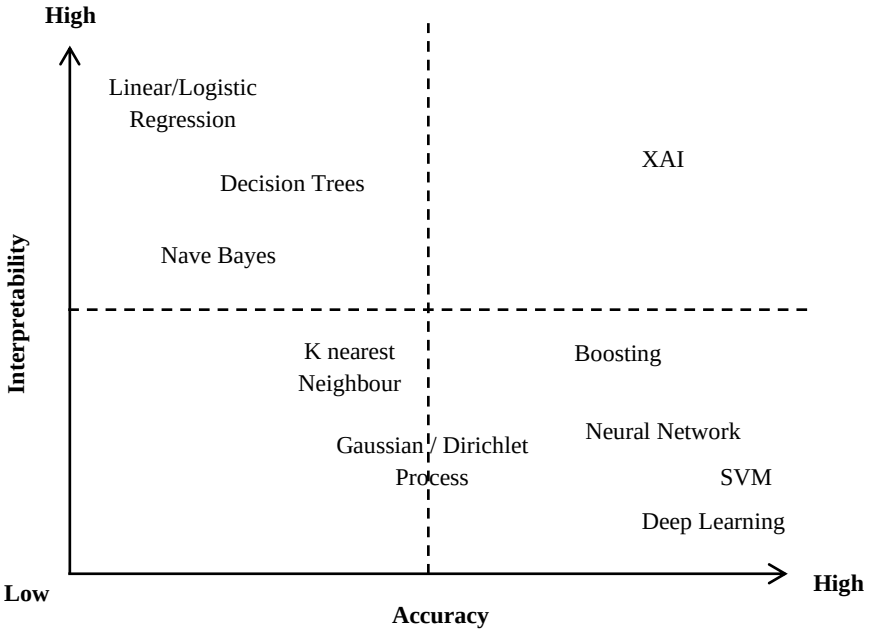

Figure 2: Accuracy VS Interpretability graph of various machine learning algorithms

2. Literature review of related work

XAI is a new emerging research field in the area of machine learning. In the past few years, there is a boom in research work related to XAI [5][7][8][9][10][11][12][13][14][15]

Researchers try to explain the working of the BBM by learning from a surrogate model (model used to explain other models) to produce the same input and output relation of a BBM which need to be explained e.g., LIME [16] and knowledge distillation [17]. Some researchers also tried to explain a model by extracting input features or intermediate generated features like (Grad-CAM) [18] and Shapley Additive exPlanations (SHAP) [19]. Most of the work in XAI is done by providing a visual explanation of the model[13] [20]or towards a specific domain like medical intelligence [21].

Mostly XAI models are data-driven models, which generate an explanation of the model purely from the dataset, without any external or prior knowledge. Researchers generally explain BBM on global or local level, based on scope of the explanation and try to explain whole working of model or just explain a particular part of the work with respect to the dataset. Interpretability of explanation model

on behalf of task and methods can be categorized in two parts; local level and global level also called explanation scope which is described below in the next section.

2.1 Global level explanation

Global explainability can be driven by extracting interpretability from BBM by training a global surrogate model or by estimating the importance of features of BBM by removing irrelevant features. The decision tree is a popular method used for the explanation of BBM. TREPEN [22] is one of the methods used to explain the working of NN using a decision tree by maximizing the ratio game [23].

The main problem with decision tree generation to explain any BBM is its length which is overcome by [24] Author presented a robust approach to generate a decision tree of small height for the explanation. In a similar way, [25]tried to create a decision tree with help of genetic programming. On other hand,[26] introduced a concept tree to build a decision tree by using expert knowledge.

The importance of identifying the main feature is also used in the global explanation of any BBM. In this technique, the model is explained by analyzing the contribution of the feature on a local or global level. Permutation Feature Importance (PFI) is the famous approach in terms of feature-based explanation presented by [27] using Random Forest. The performance of each tree (in terms of MSE) is computed with and without permuted values of interesting features. If permutation impacts a large drop in performance, then the respective feature is considered to be an important one. Later, the author of [28] extended the same idea to a model-agnostic version. It is also found that the PFI of a feature shows some unexpected effects of correlated interaction between features. [29] presented a permutation-based Shapley feature importance approach to overcome this issue.

The global features approach can also be described through visualization. In reference, [30] made Partial Dependency Plat (PDP) to analyse the impact of features in the model. This PDP idea works on displaying the non-linearity of BBM. But sometimes, it is complicated when features are highly correlated and this problem is solved by [31] by presenting the Accumulated Local Effects (ALE) new visualization model.

Instead of finding a different way to explain model, researchers redesign the model which is self-explainable. [23] presented a modified CNN model, where part templates were used as a mask on top of the conventional CNN layer. In the feature map, different activation locations are represented by

different templates. For reducing the number of templates, the input part template of the filter activates a specific location on the feature map and this is done by regularising the loss function. Same way [32] proposed a region-based interpretable CNN model in which an And-Or graph (AOG) is used for object detection. In other methods, [33] did clustering after the pooling layer to extract visual concepts.

2.2 Local level explanation:

Talking about local level explanation then LIME [16] is one of the most popular models. It uses a linear surrogate model which used data generated by BBM on a particular instance, to find the importance of features. A similar type of work is also done by ANCHOR [34], It tried to identify the most impactful area in inputs, called anchors. It used the model's input, and output mapping to explain decision boundaries on the local level. Inheriting both ideas LIME and ANCHOR, [35] present LORE (Local Rule-based Explanations). LORE used rules extracted using a decision tree to explain the local level explanation.

Apart from LIME, one more known explanation tool Saliency Maps is used by [32] for a visual explanation. Here the importance of features is shown by highlighting the aspect of the input sample as a mask to show how the model interpreted input data. The result of Saliency Maps is further improved by various methods like gradients [36], deepLIFT[37], and GradCAM [18]. Although there are many existing tools that are also used widely, DeConvNet [38] and Layer-wise Relevance Propagation (LRP) [39][40], where DeConvNet works on a pixel level to map features and LRP works on a score basis by sending a score from output layer to input layer to identify important features.

2.3 Knowledge based explanation

All methods presented in the previous section till now, are data-driven based; which means the given input dataset is used to explain the working of BBM by checking the effect of input features on output in different ways. In the past few years, a new technique known as knowledge-based explanation is very popular among researchers. Instead of using only input data, the use of some external knowledge can make the explanation more users friendly. The main objective of XAI is to explain the working of any BBM not only to the developer but to the end user also. By adding some external knowledge (familiar to end user), it is easy to understand the working of the model because the user can correlate the explanation with knowledge. This is just one example of how knowledge can be used in XAI.

Hinton [17] introduced a new concept of knowledge distillation method for model extraction. The author made a student model and trained it to mimic the complex unexplained model by optimizing a loss function. Knowledge is extracted from the network and expressed that knowledge into the hierarchal decision (a soft decision tree). To explains the working of Deep nets, the transformation of NN into a decision tree is done by training the decision tree using stochastic gradient descent with help of NN prediction. In [41] the model is extended by formulating a multi output regression problem by distillation of a vanilla tree. The same distillation method is also adopted by Tan et al. [42] to explain BBM in terms of feature shapes by extraction of iGAMs [43].

Kim et al. [44] proposed the idea of Concept Activation Vector (CAV), where researchers tried to find the importance of concept (Knowledge) in classification tasks by TCAV score. For example, the importance of strips in the classification of zebra images from others. [45] used the same TCAV for the classification of similar types of medical images based on the doctor's interest.

In [46] the author proposed a model in which humans interact with the semantic pattern from a pre train CNN model. This CNN model is then further converted into an interpretable 'And-Or graph' (AOG) model. Here humans can check model performance by guiding it with a selected feature pattern. In this type of explanation model human can introduce his knowledge to the system but in the TCAV-like model, it is still difficult to introduce knowledge if there are a large number of concepts because the problem is user has to define all types of concepts and patterns. It helps humans to incorporate with the model to check it's working. For example, in the classification of zebra images, a strip can be defined as a concept but it is not enough when the classification object contains many features. This type of problem can be solved by making a (standard and generalized) knowledge corpus.

Reference [47] explained the working of the hidden unit of the CNN model with the help of the human interaction concept. Here activation map is generated from the input and the activation region is extracted from the feature map if it has a value greater than the threshold. [48] developed a method that explains the Reinforcement learning agent's behavior in verbal communication. Here RNN is trained by corpus to explain an action taken by a model in a natural language format. The same idea is used by many researchers [49] [50][51]to explain action taken by RNN in natural language format.

In XAI, researchers also represented knowledge acknowledge graph (KG) [52][53] [54][55]and used to represent explanation in internal and external route form. Explanation model using a knowledge graph is designed considering entities, relations, paths, rules, etc in KG. This type of explanation is mainly used in the recommendation system. [56] et al. used a knowledge graph with the sequential model for a song recommendation system. The basic idea is, the extraction of different paths from a knowledge graph created by the history of user interaction, and a score is assigned to a path based on the Long Short-Term Memory model. The same type of work is represented in a different way by Ma et al. [52]. The author integrates the rule learning module with the recommendation module by extracting inductive rules between items from the knowledge graph.

Some researchers tried to embed knowledge graphs during objective function designing. [54] proposed a method, based on the given objection function. The author tried to suggest an entity corresponding to the input query and at the same time, they also extracted concepts from the knowledge graph as an explanation. In another way knowledge graph is combined with reasoning to explain any model. Generally, this type of explanation of BBM is explained in form of rules which represents some relations between input and output [57][58] or high-level concepts which represent class expression [59][60]. [61] made an explainable recommendation system that recommended treatment to a diabetic patient by combining adductive reasoning with the knowledge-based model.

Explanation of ML is also done by using rule extraction. The first attempt was made in 1994 when [62] tried to extract rules from NN Model. Rule extractions methods are mainly dependent upon the model to be explained [63], Johansson et al proposed a generalized method called G-Rex which used genetic programming to rule extraction. G-Rex was further extended by [64] who introduced decision tree and fuzzy in rule extraction. A rule extraction was also done by utilizing ACO. An ANTmnr ACO [65] is designed for the classification of the dataset. After that many improved ANTmnr algorithms are proposed for ACO rule-based classification [66][67][68]. In [69] author proposed decision tree induction to extract rules from Deep NN. [70] also worked on the same rule extraction problem by availing of the RBF network. In that paper, the dataset was trained on the RBF network. During training, they allow large overlapping among the cluster of the same classes. Initial rules are generated by considering the training output and the premises of each rule are further tuned with the help of

gradient descent. In their next paper, [71] they tried to refine the rule by a separability-correlation measure (SCM) which is used to rank attribute importance. Here RBF classifier was used to find the best subset of the attributes. After that rules are extracted by giving these retained subset attributes to the RBF network. In [72] author generated a method to extract rules from SVM by combining a prototype method with SVM. This combined model allows explainability to SVM. Here decision boundaries are determined for each class in the form of ellipsoids which are then converted into if-then rules.

Guido Bologna as a researcher worked on rule extraction for different machine learning algorithms [73][74]. In one of his latest papers [75] the author extracted rules from the deep learning model. In [76] author did sentiment analysis, where the rule extraction is done on the top most layer of NN with help of Discretized Interpretable Multi-Layer Perceptron (DIMLP). In [77] author presented work on the explanation of the NN model by modifying an existing approach of Rule Extraction by Reverse Engineering (RxREN) [78].

Table 1: Well-known XAI tools based on explanation scope of the models

S.no.	XAI Models	Level
1	LIME [16]	Local
2	ANCHORS [34]	Local
3	LRP [39][40]	Local
4	DeepTaylor [79]	Local
5	Saliency [32]	Local
6	Gradient [36]	Local
7	DeepLIFT [37]	Local
8	DeConvNet [38]	Local
9	Grad-CAM [18]	Local
10	Info Flow [80]	Global
11	TCAV [44]	Knowledge Based

From the research work available on global and local-based explanation models it is found that local explanation models are very popular. Some of the local explanation methods are also available as a tool (as shown in table 1) for the interpretation and explanation of the machine learning model.

In recent publications related to medical field, XAI is used in COVID-19 severity prediction task [81], pneumonia detection [82], pneumonia and COVID-19 detection [83], explainable ECG monitoring system [84] and glaucoma identification [85][86]. In recent trends, XAI is also used in environmental and economic fields. Reference [87] proposed an explainable model for the prediction of land sliding

and [88] proposed an XAI model to detect the wildfire from a live stream. From an economic and financial point of view reference [89] designed an XAI model for stock forecasting and [90] used XAI for the prediction of the economic growth of countries.

3. Limitation and research gap

It is found that most of the explainable models are data-driven and within the data-driven models most work is done to explain image processing-related applications (for example medical images). The visual explanation of the model is done in the form of a graph, semantic mask, and saliency map, and most of the visual explanations are limited to the heat map.

As stated in the previous section, knowledge base explanation is a newly emerging area in XAI that holds a lot of potentials to be explored further. In this synopsis, a framework is proposed to extend this new emerging knowledge-based explanation. In these knowledge-based explanations, either knowledge is extracted from the model in form of an explanation or some external knowledge is added to the model which helps in the explanation. There are some open questions, one of which is how to identify or retrieve required useful knowledge from a vast knowledge set. Considering these limitations and questions in mind, the following research framework is designed.

4. Framework of the proposed research

The research objectives based on the research gap identified in section 3 are as follows:

- Systematic and comparative study of various knowledge-based approaches for the explainable model as mentioned in section 3.
- The design of a knowledge-based model for explaining a simple neural network model can be further extended to a deep learning model.
- Application and analysis of developed explainable model on various benchmark and real-world applications.
- Development of software tool for knowledge-based explainable intelligent systems.

It is proposed to limit our research in this thesis to an appropriate mixture and subset of the issues identified above.

5. Methodology

In the proposed method, knowledge will be used to the explanation of the machine learning model. As said in the previous section this knowledge can be exploited in two ways:

Firstly, extracting knowledge from the target BBM and representing this knowledge in the form of rules as an explanation. To extract the knowledge methods like fuzzy system, Decision tree, or knowledge graph will be explored.

The second way, instead of extracting knowledge, additional knowledge (like in the TCAV [Kim] model) can be used to explain the model. Instead of using these two models separately, a hybrid model can be generated by extracting an intermediate knowledge from (target BBM) system and this intermediate knowledge can be used with other models to explain the working of our target BBM.

Explainable AI is used in various applications domain especially in the medical field [91][92][93]. In other applications [94] designed an XAI model for intrusion detection systems, and [95] designed an interpretable model for smart manufacturing.

The XAI can also be used in other application domains like disease prediction, stock prediction, explanation of recommendation system, online shopping items recommendation, recommendation of study material, or any similar types of application.

Based on the above-proposed framework stepwise methodologies are as follows:

- ***Exploring existing models:*** Existing models related to knowledge-based XAI will be studied. A comparative study will be done on behalf of the model's methodology, results, applications, and limitations.
- ***Model development:*** A new model will be designed based on knowledge using Python programming language. In case of a requirement of GUI, Django (Python-based web framework) will be used to implement it. In the proposed model the knowledge can be an additional feature introduced to the model or any intermediate data generated by the model. This knowledge will be extracted using a fuzzy, decision tree, knowledge graph, etc. After completion of the model, the performance of the model will also be analysed on different parameters.
- ***Application*****:** The developed model will be used in real-world applications like disease prediction, items recommendation for online shopping, etc.

- ***Tool development*:** Finally, a software tool will also be developed to explore the working of the proposed model. With this tool, users could be able to check the working of the model and analyse it using different types of graphs and tables.

The tentative timeline of work will be as follows. The study of various knowledge-based approaches along with preliminary experiments on benchmark datasets will be done in the first year. It will be followed by the design and development of the proposed model and its implementation in the second year, following which the experimentation and thesis writing will continue in the third year.

References:

[1] T. Revell, "Computer says 'no comment,'" New Scientist, vol. 238, no. 3173, pp. 40–43, Apr. 2018, doi: 10.1016/S0262-4079(18)30664-X.

[2] D. Silver et al., "Mastering the game of Go with deep neural networks and tree search," Nature 2016 529:7587, vol. 529, no. 7587, pp. 484–489, Jan. 2016, doi: 10.1038/ nature16961.

[3] Alex Shandro, "House of Lords - AI in the UK: ready, willing and able? - Artificial Intelligence Committee."https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10014.htm

[4] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision making and a 'right to explanation,'" AI Magazine, vol. 38, no. 3, pp. 50–57, Sep. 2017, doi: 10.1609/AIMAG.V38I3.2741.

[5] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," AI Magazine, vol. 40, no. 2, pp. 44–58, Jun. 2019, doi: 10.1609/AIMAG.V40I2.2850.

[6] Fan Yang, "Evaluating Explanation Without Ground Truth in Interpretable Machine Learning," arXiv:1907.06831, 2019. https://arxiv.org/abs/1907.06831.

[7] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, Sep. 2018, doi: 10.1109/ ACCESS.2018.2870052.

[8] F. K. Dosilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," 2018 41st International Convention on Information and Communication Technology, Electronics

and Microelectronics, MIPRO 2018 - Proceedings, pp. 210–215, Jun. 2018, doi: 10.23919/MIPRO.2018.8400040.

[9] C. Molnar, G. Casalicchio, and B. Bischl, “Interpretable Machine Learning -- A Brief History, State-of-the-Art and Challenges,” Communications in Computer and Information Science, vol. 1323, pp. 417–431, Oct. 2020, https://arxiv.org/abs/2010.09337v1

[10] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning,” Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018, pp. 80–89, Jan. 2019, doi: 10.1109/DSAA.2018.00018.

[11] DuMengnan, LiuNinghao, and HuXia, “Techniques for interpretable machine learning,” Commun ACM, vol. 63, no. 1, pp. 68–77, Dec. 2019, doi: 10.1145/3359786.

[12] GuidottiRiccardo, MonrealeAnna, RuggieriSalvatore, TuriniFranco, GiannottiFosca, and PedreschiDino, “A Survey of Methods for Explaining Black Box Models,” ACM Computing Surveys (CSUR), Aug. 2018, doi: 10.1145/3236009.

[13] O. Biran and C. v. Cotton, “Explanation and Justification in Machine Learning : A Survey Or,” IJCAI, Workshop on Explainable Artificial Intelligence (XAI), 2017.

[14] F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning,” arXiv: Machine Learning, 2017.

[15] Dr. Freddy Lecue, “XAI – The story so far,” The Montreal Chapter of the IEEE Systems, 2019. https://montreal.ieee.ca/event/xai-the-story-so-far.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should i trust you?’ Explaining the predictions of any classifier,” Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 13-17-August-2016, pp. 1135–1144, Aug. 2016, doi: 10.1145/2939672.2939778.

[17] G. E. Hinton, O. Vinyals, and J. Dean, “Distilling the Knowledge in a Neural Network,” ArXiv, 2015.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” Proceedings of

the IEEE International Conference on Computer Vision, vol. 2017-October, pp. 618–626, Dec. 2017, doi: 10.1109/ICCV.2017.74.

[19] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” Advances in Neural Information Processing Systems, vol. 30, 2017, https://github.com/slundberg/shap

[20] Y. Ming, “A Survey on Visualization For Explainable Classifiers,” 2017, https://www.semanticscholar.org/paper/A-SURVEY-ON-VISUALIZATION-FOR-EXPLAINABLE -by-Ming/b6f31303c03684b92ca7e44691fc1155a492a15c.

[21] E. Tjoa and C. Guan, “A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI,” IEEE Transactions on Neural Networks and Learning Systems, pp. 1–21, Oct. 2020, doi: 10.1109/TNNLS.2020.3027314.

[22] Mark W. Craven and Jude W. Shavlik, “Extracting tree-structured representations of trained networks,” NIPS, 1995. https://dl.acm.org/doi/10.5555/2998828.2998832.

[23] Q. Zhang, Y. N. Wu, and S.-C. Zhu, “Interpretable Convolutional Neural Networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8827–8836.

[24] R. Krishnan, G. Sivakumar, and P. Bhattacharya, “Extracting decision trees from trained neural networks,” Pattern Recognition, vol. 32, no. 12, pp. 1999–2009, Dec. 1999, doi: 10.1016/S0031-3203(98)00181-2.

[25] U. Johansson and L. Niklasson, “Evolving decision trees using oracle guides,” 2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009 - Proceedings, pp. 238–244, 2009, doi: 10.1109/CIDM.2009.4938655.

[26] X. Renard, N. Woloszko, J. Aigrain, and M. Detyniecki, “Concept Tree: High-Level Representation of Variables for More Interpretable Surrogate Decision Trees,” AXA OECD , 2019.

[27] L. Breiman, “Random Forests,” Machine Learning 2001 45:1, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[28] A. Fisher, C. Rudin, and F. Dominici, “All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously,” Journal of Machine Learning Research, vol. 20, no. 177, pp. 1–81, 2019, http://jmlr.org/papers/v20/18-760.html

[29] G. Casalicchio, C. Molnar, and B. Bischl, “Visualizing the feature importance for black box models,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11051 LNAI, pp. 655–670, 2019, doi: 10.1007/978-3-030-10925-7_40.

[30] J. H. Friedman, “Greedy function approximation: A gradient boosting machine.,” https://doi.org/10.1214/aos/1013203451, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/AOS/1013203451.

[31] D. W. Apley and J. Zhu, “Visualizing the effects of predictor variables in black box supervised learning models,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 82, no. 4, pp. 1059–1086, Sep. 2020, doi: 10.1111/RSSB.12377.

[32] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps,” CoRR, 2014.

[33] J. Wang et al., “Visual concepts and compositional voting,” Annals of Mathematical Sciences and Applications, vol. 3, no. 1, pp. 151–188, 2018, doi: 10.4310/AMSA.2018.V3.N1.A5.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin, “Anchors: High-Precision Model-Agnostic Explanations,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, Apr. 2018, https://ojs.aaai.org/index.php/AAAI/article/view/11491

[35] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, “Local Rule-Based Explanations of Black Box Decision Systems,” arXiv:1805.10820, May 2018, http://arxiv.org/abs/1805.10820

[36] M. Alber et al., “iNNvestigate neural networks!,” Journal of Machine Learning Research, vol. 20, Aug. 2018, https://arxiv.org/abs/ 1808.04260v1

[37] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning Important Features Through Propagating Activation Differences,” 34th International Conference on Machine Learning, ICML 2017, vol. 7, pp. 4844–4866, Apr. 2017, https://arxiv.org/abs/1704.02685v2

[38] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, “Deconvolutional networks,” Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2528–2535, 2010, doi: 10.1109/CVPR.2010.5539957.

[39] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” PLoS ONE, vol. 10, no. 7, Jul. 2015, doi: 10.1371/JOURNAL.PONE.0130140.

[40] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, “Towards better understanding of gradient-based attribution methods for Deep Neural Networks,” ICLR 2018 Conference. Feb. 15, 2018.

[41] X. Liu, X. Wang, and S. Matwin, “Improving the Interpretability of Deep Neural Networks with Knowledge Distillation,” IEEE International Conference on Data Mining Workshops, ICDMW, vol. 2018-November, pp. 905–912, Dec. 2018, https://arxiv.org/abs/ 1812.10924v1

[42] S. Tan, R. Caruana, G. Hooker, P. Koch, and A. Gordo, “Learning Global Additive Explanations for Neural Nets Using Model Distillation,” arXiv: Machine Learning, 2018.

[43] R. C. S. Tan, “Detecting Bias in Black-Box Models Using Transparent Model Distillation. - AMiner,” AIES, 2017. https://www.aminer.cn/pub/5a260c8117c44a4ba8a31077/detecting-bias-in-black-box-models-using-transparent-model-distillation.

[44] B. Kim et al., “Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV),” 35th International Conference on Machine Learning, ICML 2018, vol. 6, pp. 4186–4195, Nov. 2017, https://arxiv.org/abs/1711.11279v5

[45] C. J. Cai et al., “Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making,” Conference on Human Factors in Computing Systems - Proceedings, Feb. 2019, https://arxiv.org/abs/1902.02960v1

[46] Q. Zhang, R. Cao, S. Zhang, M. Edmonds, Y. Wu, and S.-C. Zhu, “Interactively Transferring CNN Patterns for Part Localization,” ArXiv, 2017.

[47] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, “Network dissection: Quantifying interpretability of deep visual representations,” Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017-January, pp. 3319–3327, Nov. 2017, doi: 10.1109/CVPR.2017.354.

[48] U. Ehsan, B. Harrison, L. Chan, and M. O. Riedl, “Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations,” AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 81–87, Dec. 2018, doi: 10.1145/3278721.3278736.

[49] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, “Automated rationale generation: A technique for explainable AI and its effects on human perceptions,” International Conference on Intelligent User Interfaces, Proceedings IUI, vol. Part F147615, pp. 263–274, 2019, doi: 10.1145/3301275.3302316.

[50] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, “Learning to Generate Natural Language Rationales for Game Playing Agents,” AIIDE Workshops, 2018.

[51] Brent Harrison, Upol Ehsan, and Mark O. Riedl, “Guiding Reinforcement Learning Exploration Using Natural Language,” AAMAS, 2018. https://dl.acm.org/doi/10.5555/3237383. 3238036.

[52] W. Ma et al., “Jointly Learning Explainable Rules for Recommendation with Knowledge Graph,” The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019, pp. 1210–1221, Mar. 2019, https://arxiv.org/abs/1903.03714v1

[53] Q. Ai, V. Azizi, X. Chen, and Y. Zhang, “Learning Heterogeneous Knowledge Base Embeddings for Explainable Recommendation,” Algorithms, vol. 11, no. 9, May 2018, doi: 10.3390/a11090137.

[54] Y. Zhang, Y. Xiao, S. W. Hwang, H. Wang, X. S. Wang, and W. Wang, “Entity suggestion with conceptual explanation,” IJCAI International Joint Conference on Artificial Intelligence, pp. 4244–4250, 2017, doi: 10.24963/IJCAI.2017/593.

[55] J. Huang, W. X. Zhao, H. Dou, J. R. Wen, and E. Y. Chang, “Improving sequential recommendation with knowledge-enhanced memory networks,” 41st International ACM

SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018, pp. 505–514, Jun. 2018, doi: 10.1145/3209978.3210017.

[56] X. Wang, D. Wang, C. Xu, X. He, Y. Cao, and T.-S. Chua, "Explainable Reasoning over Knowledge Graphs for Recommendation," 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, pp. 5329–5336, Nov. 2018, https://arxiv.org/abs/1811.04540v1

[57] M. Labaf, P. Hitzler, and A. B. Evans, "Propositional Rule Extraction from Neural Networks under Background Knowledge," NeSy, 2017.

[58] G. Paçacı, D. Johnson, S. McKeever, and A. Hamfelt, "Explaining black box models with inductive synthesis," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11540 LNCS, pp. 334–345, 2019, doi: 10.1007/978-3-030-22750-0_27.

[59] M. K. Sarker and P. Hitzler, "Efficient Concept Induction for Description Logics," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 3036–3043, Jul. 2019, doi: 10.1609/AAAI.V33I01.33013036.

[60] M. K. Sarker, N. Xie, D. Doran, M. Raymer, and P. Hitzler, "Explaining Trained Neural Networks with Semantic Web Technologies: First Steps," CEUR Workshop Proceedings, vol. 2003, Oct. 2017, https://arxiv.org/abs/ 1710.04324v1

[61] Oshani Seneviratne et al., "Events / Enabling Trust in Clinical Decision Support Recommendations through Semantics," 2019. https://events.unimelb.edu.au/event/6708-enabling-trust-in-clinical-decision-support.

[62] M. W. Craven and J. W. Shavlik, "Using Sampling and Queries to Extract Rules from Trained Neural Networks," Machine Learning Proceedings 1994, pp. 37–45, Jan. 1994, doi: 10.1016/B978-1-55860-335-6.50013-1.

[63] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," Knowledge-Based Systems, vol. 8, no. 6, pp. 373–389, Dec. 1995, doi: 10.1016/0950-7051(96)81920-4.

[64] U. Johansson, R. König, and L. Niklasson, “The Truth is In There - Rule Extraction from Opaque Models Using Genetic Programming,” FLAIRS Conference, 2004.

[65] R. S. Parpinelli, H. S. Lopes, and A. A. Freitas, “Data mining with an ant colony optimization algorithm,” IEEE Transactions on Evolutionary Computation, vol. 6, no. 4, pp. 321–332, Aug. 2002, doi: 10.1109/TEVC.2002.802452.

[66] F. E. B. Otero, A. A. Freitas, and C. G. Johnson, “A new sequential covering strategy for inducing classification rules with ant colony algorithms,” IEEE Transactions on Evolutionary Computation, vol. 17, no. 1, pp. 64–76, 2013, doi: 10.1109/TEVC.2012.2185846.

[67] J. L. Olmo, J. R. Romero, and S. Ventura, “Using ant programming guided by grammar for building rule-based classifiers,” IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 41, no. 6, pp. 1585–1599, Dec. 2011, doi: 10.1109/TSMCB.2011.2157681.

[68] F. E. B. Otero and A. A. Freitas, “Improving the interpretability of classification rules discovered by an ant colony algorithm: Extended results,” Evolutionary Computation, vol. 24, no. 3, pp. 385–409, Sep. 2016, doi: 10.1162/EVCO_A_00155.

[69] J. R. Zilke, E. Loza Mencía, and F. Janssen, “DeepRED – Rule Extraction from Deep Neural Networks,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9956 LNAI, pp. 457–473, Oct. 2016, doi: 10.1007/978-3-319-46307-0_29.

[70] L. Wang and X. Fu, “A Simple Rule Extraction Method Using a Compact RBF Neural Network,” Lecture Notes in Computer Science, vol. 3496, no. I, pp. 682–687, 2005, doi: 10.1007/11427391_109.

[71] X. Fu and L. Wang, “Extracting Rules from RBF Neural Networks Using Reduced Data Attribute Sets,” 2014.

[72] D. Martens, J. Huysmans, R. Setiono, J. Vanthienen, and B. Baesens, “Rule extraction from support vector machines: An overview of issues and application in credit scoring,” Studies in Computational Intelligence, vol. 80, pp. 33–63, 2008, doi: 10.1007/978-3-540-75390-2_2.

[73] G. Bologna and Y. Hayashi, “Characterization of symbolic rules embedded in deep DIMLP networks: A challenge to transparency of deep learning,” Journal of Artificial Intelligence and

Soft Computing Research, vol. 7, no. 4, pp. 265–286, Oct. 2017, doi: 10.1515/JAISCR-2017-0019.

[74] G. Bologna and C. Pellegrini, “Constraining the MLP power of expression to facilitate symbolic rule extraction,” IEEE International Conference on Neural Networks - Conference Proceedings, vol. 1, pp. 146–151, 1998, doi: 10.1109/IJCNN.1998.682252.

[75] G. Bologna and Y. Hayashi, “A rule extraction study on a neural network trained by deep learning,” Proceedings of the International Joint Conference on Neural Networks, vol. 2016-October, pp. 668–675, Oct. 2016, doi: 10.1109/IJCNN.2016.7727264.

[76] G. Bologna, “A Rule Extraction Study Based on a Convolutional Neural Network,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11015 LNCS, pp. 304–313, Aug. 2018, doi: 10.1007/978-3-319-99740-7_22.

[77] S. K. Biswas, M. Chakraborty, B. Purkayastha, P. Roy, and D. M. Thounaojam, “Rule Extraction from Training Data Using Neural Network,” https://doi.org/10.1142/S0218213017500063, vol. 26, no. 3, Dec. 2016.

[78] M. G. Augasta and T. Kathirvalavakumar, “Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems,” Neural Processing Letters 2011 35:2, vol. 35, no. 2, pp. 131–150, Dec. 2011, doi: 10.1007/S11063-011-9207-8.

[79] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K. R. Müller, “Explaining nonlinear classification decisions with deep Taylor decomposition,” Pattern Recognition, vol. 65, pp. 211–222, May 2017, doi: 10.1016/J.PATCOG.2016.11.008.

[80] K. Wongsuphasawat et al., “Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow,” IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pp. 1–12, Jan. 2018, doi: 10.1109/TVCG.2017.2744878.

[81] H. Wu et al., “Interpretable Machine Learning for COVID-19: An Empirical Study on Severity Prediction Task,” IEEE Transactions on Artificial Intelligence, pp. 1–1, Jun. 2021, doi: 10.1109/TAI.2021.3092698.

[82] H. Ren et al., “Interpretable Pneumonia Detection by Combining Deep Learning and Explainable Models with Multisource Data,” IEEE Access, vol. 9, pp. 95872–95883, 2021, doi: 10.1109/ACCESS.2021.3090215.

[83] L. Zou et al., “Ensemble image explainable AI (XAI) algorithm for severe community-acquired pneumonia and COVID-19 respiratory infections,” IEEE Transactions on Artificial Intelligence, pp. 1–1, Feb. 2022, doi: 10.1109/TAI.2022.3153754.

[84] A. Raza, K. P. Tran, L. Koehl, and S. Li, “Designing ECG monitoring healthcare system with federated transfer learning and explainable AI,” Knowledge-Based Systems, vol. 236, p. 107763, Jan. 2022, doi: 10.1016/J.KNOSYS.2021.107763.

[85] S. Oh, Y. Park, K. J. Cho, and S. J. Kim, “Explainable Machine Learning Model for Glaucoma Diagnosis and Its Interpretation,” Diagnostics 2021, Vol. 11, Page 510, vol. 11, no. 3, p. 510, Mar. 2021, doi: 10.3390/DIAGNOSTICS11030510.

[86] Md. S. Kamal, N. Dey, L. Chowdhury, S. I. Hasan, and K. Santosh, “Explainable AI for Glaucoma Prediction Analysis to Understand Risk Factors in Treatment Planning,” IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1–9, 2022, doi: 10.1109/TIM.2022.3171613.

[87] E. Collini, L. A. I. Palesi, P. Nesi, G. Pantaleo, N. Nocentini, and A. Rosi, “Predicting and Understanding Landslide Events with Explainable AI,” IEEE Access, vol. 10, pp. 31175–31189, 2022, doi: 10.1109/ACCESS.2022.3158328.

[88] T. C. Phan, N. D. K. Quach, T. T. Nguyen, T. T. Nguyen, J. Jo, and Q. V. H. Nguyen, “Real-time wildfire detection with semantic explanations,” Expert Systems with Applications, vol. 201, p. 117007, Sep. 2022, doi: 10.1016/J.ESWA.2022.117007.

[89] S. M. Carta, S. Consoli, L. Piras, A. S. Podda, and D. R. Recupero, “Explainable Machine Learning Exploiting News and Domain-Specific Lexicon for Stock Market Forecasting,” IEEE Access, vol. 9, pp. 30193–30205, 2021, doi: 10.1109/ACCESS.2021.3059960.

[90] S. Park and J.-S. Yang, “Interpretable deep learning LSTM model for intelligent economic decision-making,” Knowledge-Based Systems, vol. 248, p. 108907, Jul. 2022, doi: 10.1016/J.KNOSYS.2022.108907.

[91] E. Tjoa and C. Guan, “A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI,” IEEE Transactions on Neural Networks and Learning Systems, pp. 1–21, Oct. 2020, doi: 10.1109/TNNLS.2020.3027314.

[92] G. Yang, Q. Ye, and J. Xia, “Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond,” Information Fusion, vol. 77, pp. 29–52, Jan. 2022, doi: 10.1016/J.INFFUS.2021.07.016.

[93] L. A. de Souza et al., “Convolutional Neural Networks for the evaluation of cancer in Barrett’s esophagus: Explainable AI to lighten up the black-box,” Computers in Biology and Medicine, vol. 135, p. 104578, Aug. 2021, doi: 10.1016/ J.COMPBIOMED. 2021.104578.

[94] M. Wang, K. Zheng, Y. Yang, and X. Wang, “An Explainable Machine Learning Framework for Intrusion Detection Systems,” IEEE Access, vol. 8, pp. 73127–73141, 2020, doi: 10.1109/ACCESS.2020.2988359.

[95] J. Zhang and R. X. Gao, “Deep Learning-Driven Data Curation and Model Interpretation for Smart Manufacturing,” Chinese Journal of Mechanical Engineering 2021 34:1, vol. 34, no. 1, pp. 1–21, Jul. 2021, doi: 10.1186/S10033-021-00587-y